Hotspot Analysis with QGIS

Demo Exercise

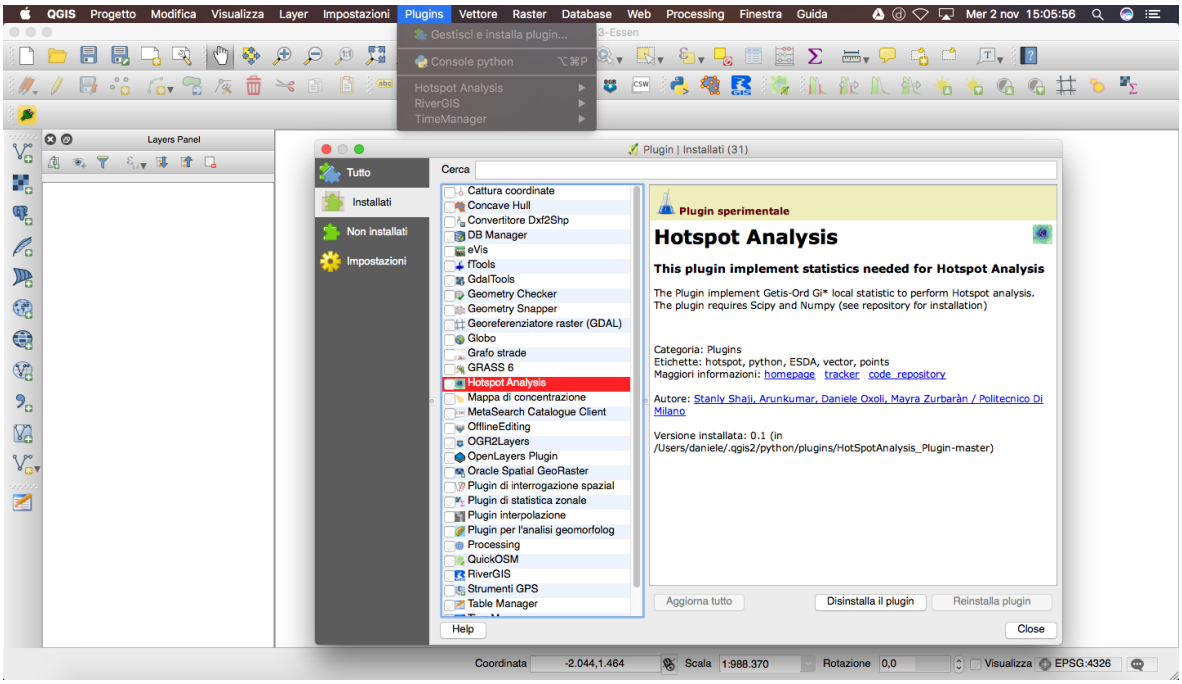
The following demo includes two examples of the QGIS Hotspot Analysis plugin application.

Plugin update: 02/11/2016

Content:

0. System check

After the installation of both the plugin and plugin's dependencies:
(see: https://github.com/stanly3690/HotSpotAnalysis_Plugin/blob/master/README.md)
Open QGIS and check if the plugin is correctly installed:

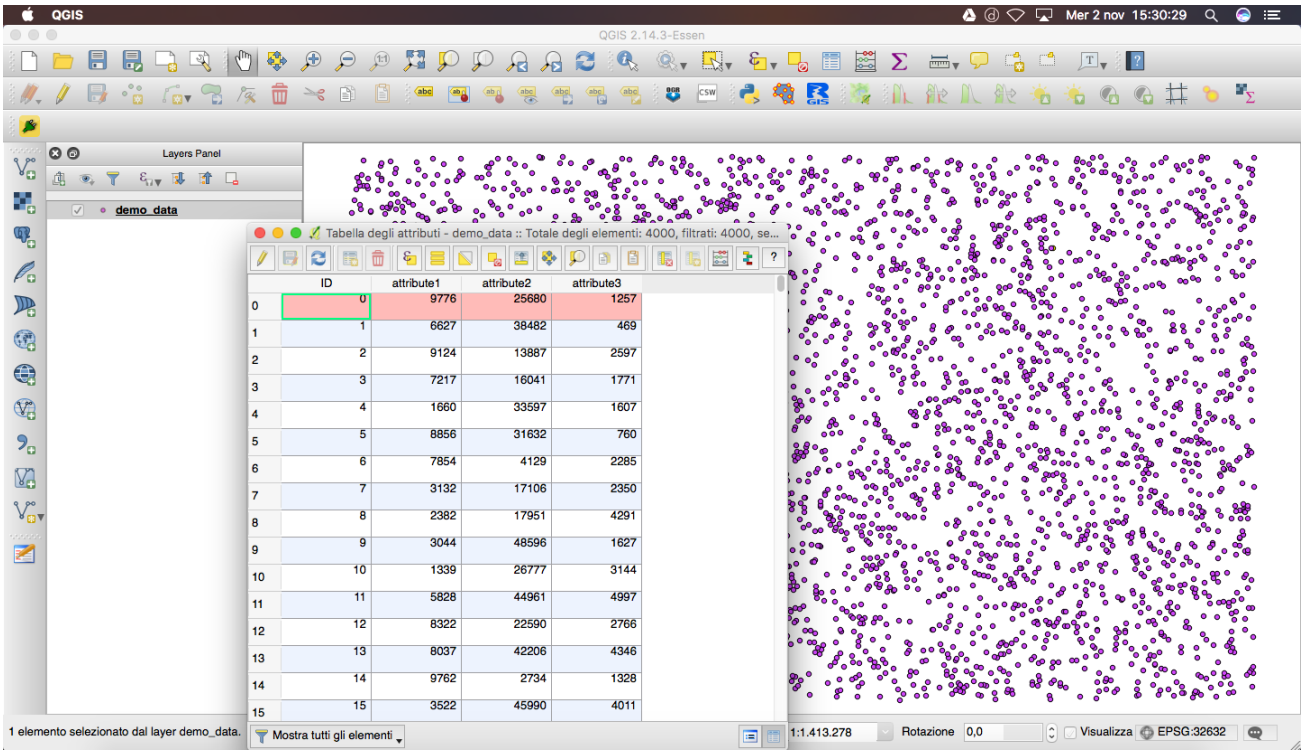


No Python errors should appear while opening QGIS or installing the plugin. If this is not true, please check that the installation procedure was performed correctly.

(Sometimes, due to the system Python Path some unexpected error may occur. If these are not solved by repeating the installation procedure you can try to uninstall QGIS, install it again and proceed with the plugin installation again)

1. Create a valid input layer

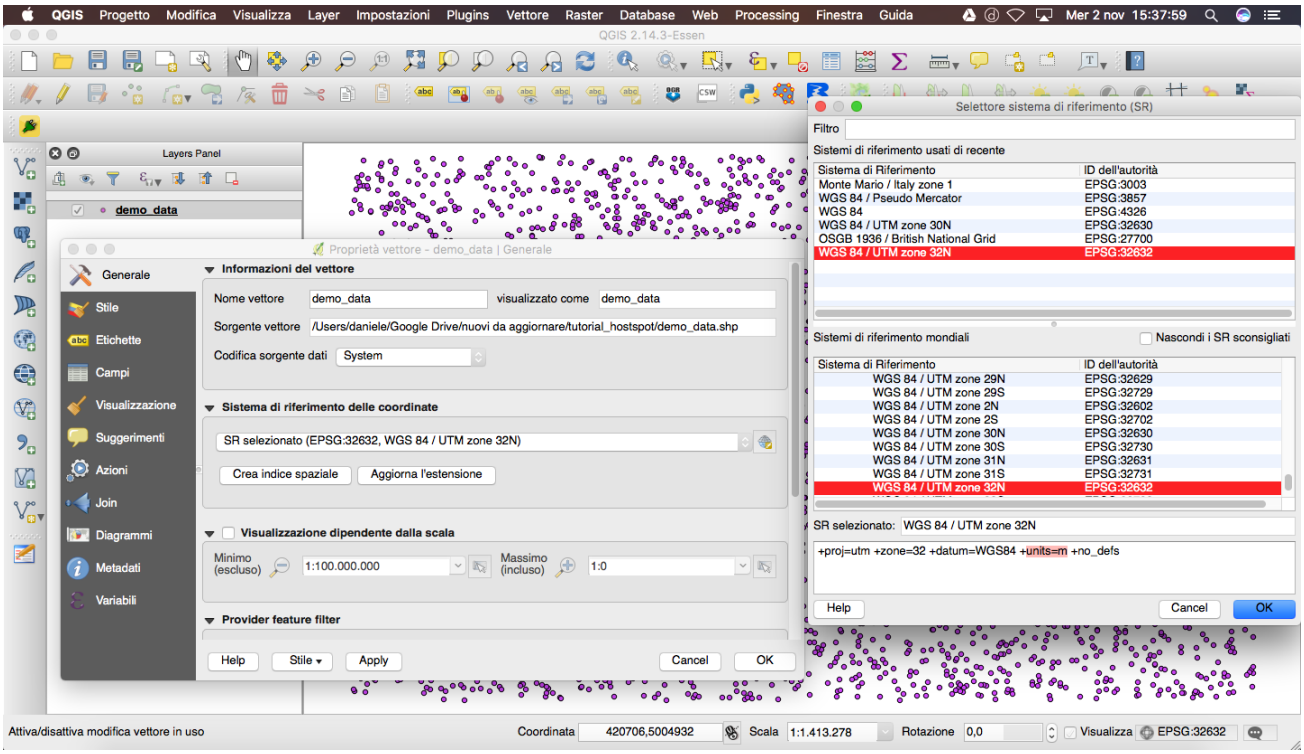
The plugin accepts as input only point shapefile with at least a positive numeric attribute assigned to any point of the dataset. A demo input shapefile is available here: https://github.com/stanly3690/HotSpotAnalysis_Plugin/tree/master/test_data - > **demo_data**
Open the **demo_data** shapefile and look at its attribute table:



As it can be seen, to any of the 4000 points of **demo_data** are assigned three positive numeric attributes. Your dataset should look like this latter, whit at least one positive numeric attribute assigned. To properly run Hotspot Analysis your dataset should contain at least ~ 30 points (indicatively).
Another requirement is the coordinate system associated to your input shapefile. The plugin requires your layer to be **projected**. Therefore, be sure about that by checking at the **layer properties**. Moreover, be aware about the **unit of measure** in which your selected projected coordinate system is expressed. With respect to the **demo_data** (see picture below), the assigned projected coordinate system is WGS84/UTM32N which is expressed in **meters**. The unit of measure is of primary importance to initialise the plugin running parameters (i.e. Distance Band).

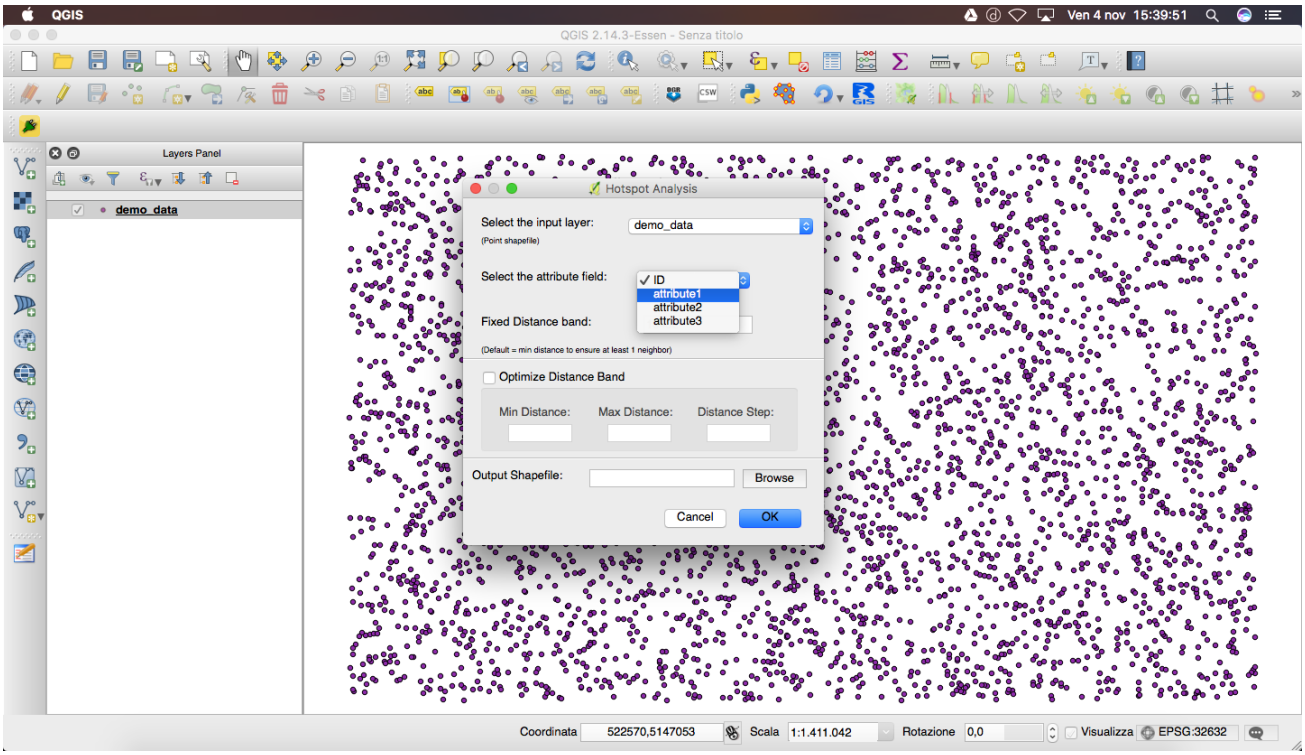
The plugin process consists of the computation of the Getis-Ord Gi local statistics, which aims to detected atypical locations (i.e. hotspots/coldspots) in the spatial arrangement of a given variable (i.e. the positive numeric attribute). Practically speaking, Gi* compare local averages with global average to underline the presence of significant high-values (or low-values) clusters. Local averages*

are computed by considering for any point of the dataset a set of neighbourhood points within a specific distance from the focal position. For this reason, the plugin requires to specify a Distance Band by using the same unit of measure of the projected coordinate system of the input shapefile.



Further information about the Distance Band have been reported in section 2.

Once these features have been checked, it possible to launch the Hotspot Analysis plugin:



When the User Interface is open, select the input layer from the list as well as specify which attribute field contains the positive numerical attribute you want to use to run Hotspot Analysis. In the case of **demo_data** select one among: "*attribute*1", "*attribute2*", "*attribute3*".

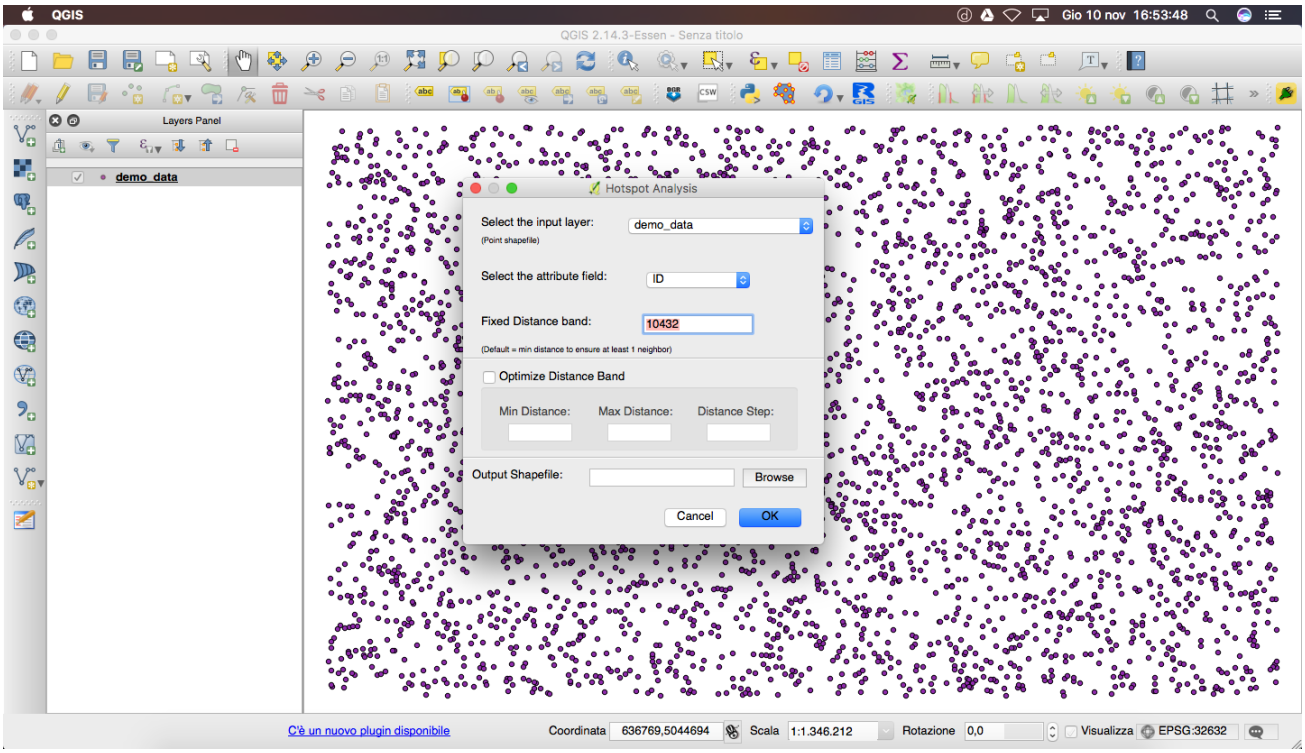
(Note: **demo_data** is an invented dataset, attributes as well as point locations do not represent any physical feature)

2. Select a proper Distance Band for the analysis

As mentioned in the previous section, in order to compute local Getis-Ord Gi* statistic it is required to specify a spatial relation between points in the dataset. The plugin allows to modelled this relation using a **Fixed Distance Band**. This means that the local statistic is computed for any point by considering a subset of neighbour points within a fixed distance.

The Distance Band selection is a crucial step but no fixed rules are available to perform this task. Analysis distance depends directly on which phenomena your dataset describes as well as on the results you are looking for (example: if you are looking for hotspots of crime cases per city block within a city, your distance may be in the order of the average distance between two neighbour city blocks). Nevertheless, you should select a distance that guarantees some neighbours to any point of your dataset in order to avoid so called "islands".

The default suggested value for the Fixed Distance Band is the **minimum distance to guarantee at least one neighbour** to any point of the dataset and it is derived from considering the point spatial distribution.
You can change this value according to you needs but keep in mind that you should not select a lower distance.

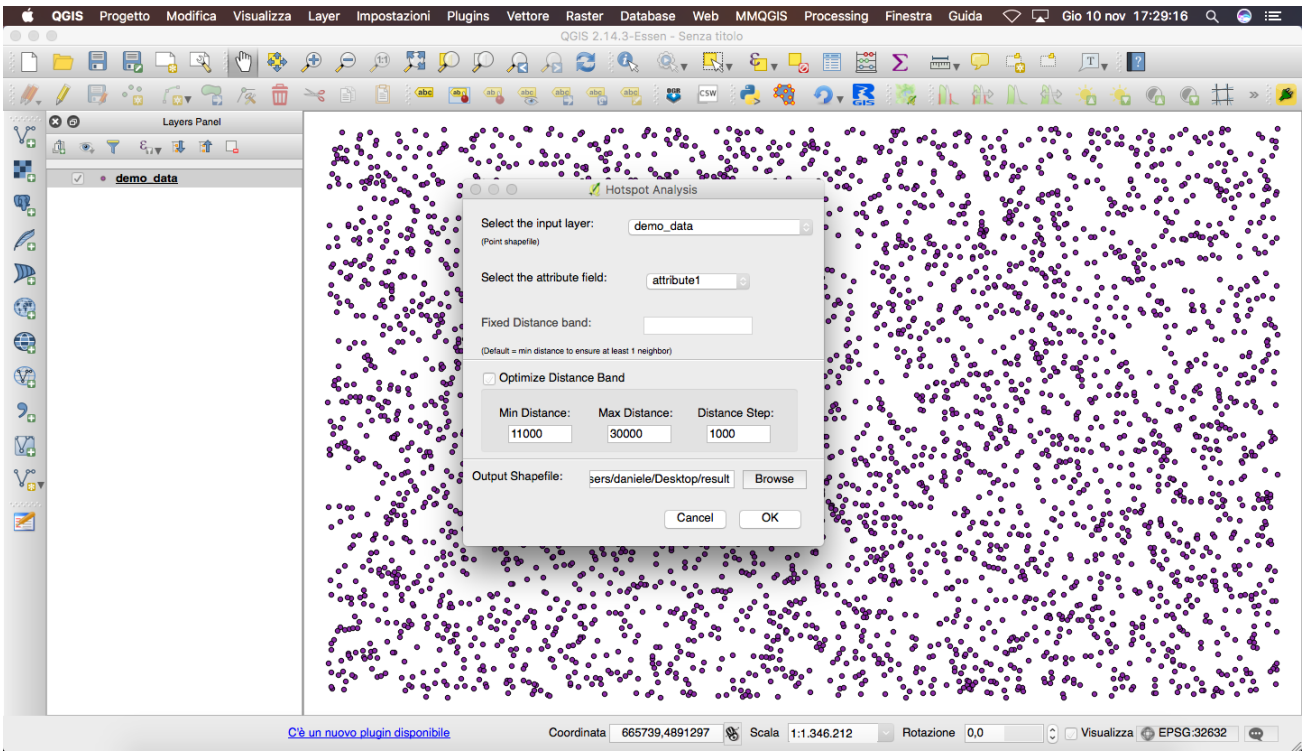


In the case of **demo_data**, default suggested value for Fixed Distance Band is = 10423 m (according to the spatial distribution of points and the projected coordinate system of the layer, which is expressed in meters by definition).

The plugin includes also an optimization procedure in order to selected the distance band. To use this option, you have to activate the command **Optimize Distance Band** by ticking the checkbox. Then you have to specify a distance range (by typing a minimum and maximum distance value) and a distance step (example: **Min Distance** = 10000 m, **Max Distance**= 30000 m, **Distance Step**= 1000 m -> the algorithm will test this array of distance: [10000 m, 11000 m, …, 29000 m, 30000 m])

The optimized distance which will be automatically used to perform Hotspot Analysis is the one that maximize the Z-score of the **global Moran's I** index for your dataset.
This index suggests at which analysis distance the dataset shows high cluster activity (either of negative or positive values) and therefore is not only based on the point spatial distribution but also on the spatial arrangement of the numeric attribute that you select for the analysis.
Be aware that the suggested distance may not agree with your analysis purposes. You should specify a suitable range of distances which fits your analysis needs as well as you should not specify a Min Distance lower that the default suggested value described in the previous section (not mandatory).

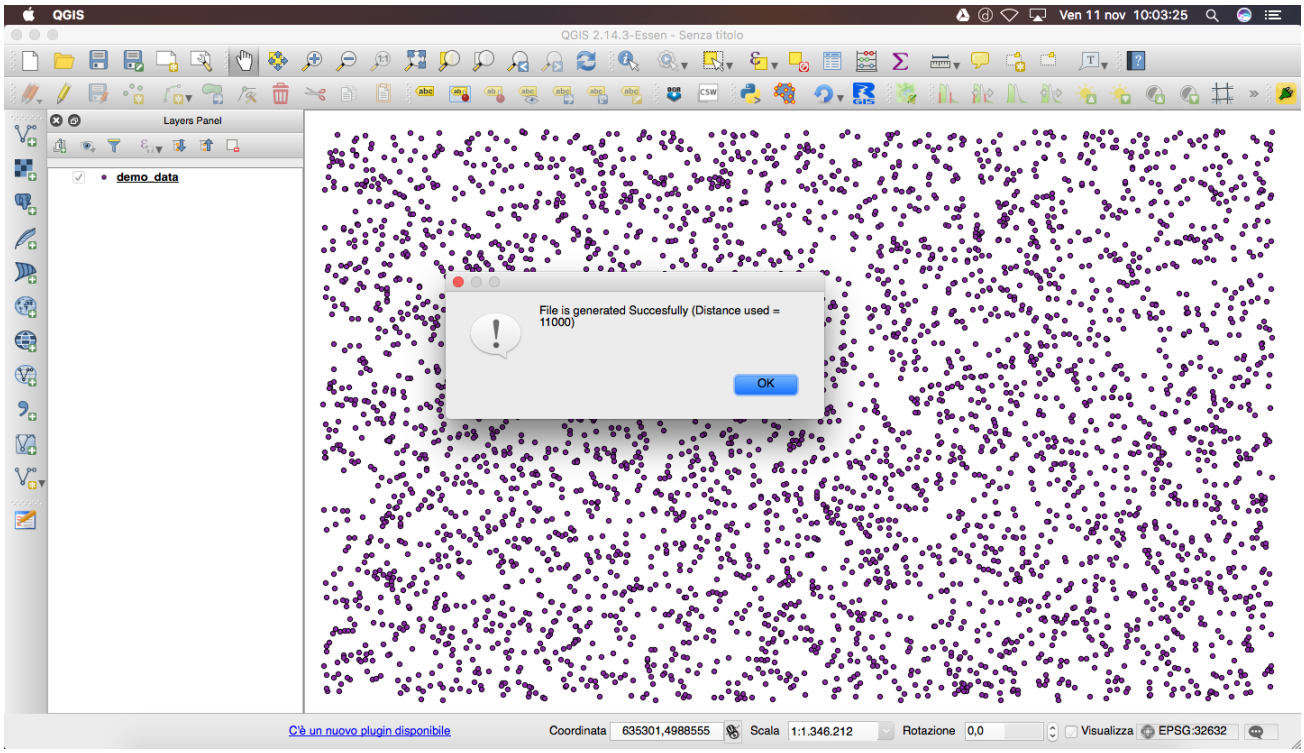
Computational time for the optimization increase according to the amount of points in your dataset but also is strongly related to the distance range and step you decide to adopt (i.e. **large distance range + little distance step = long computational time**!).

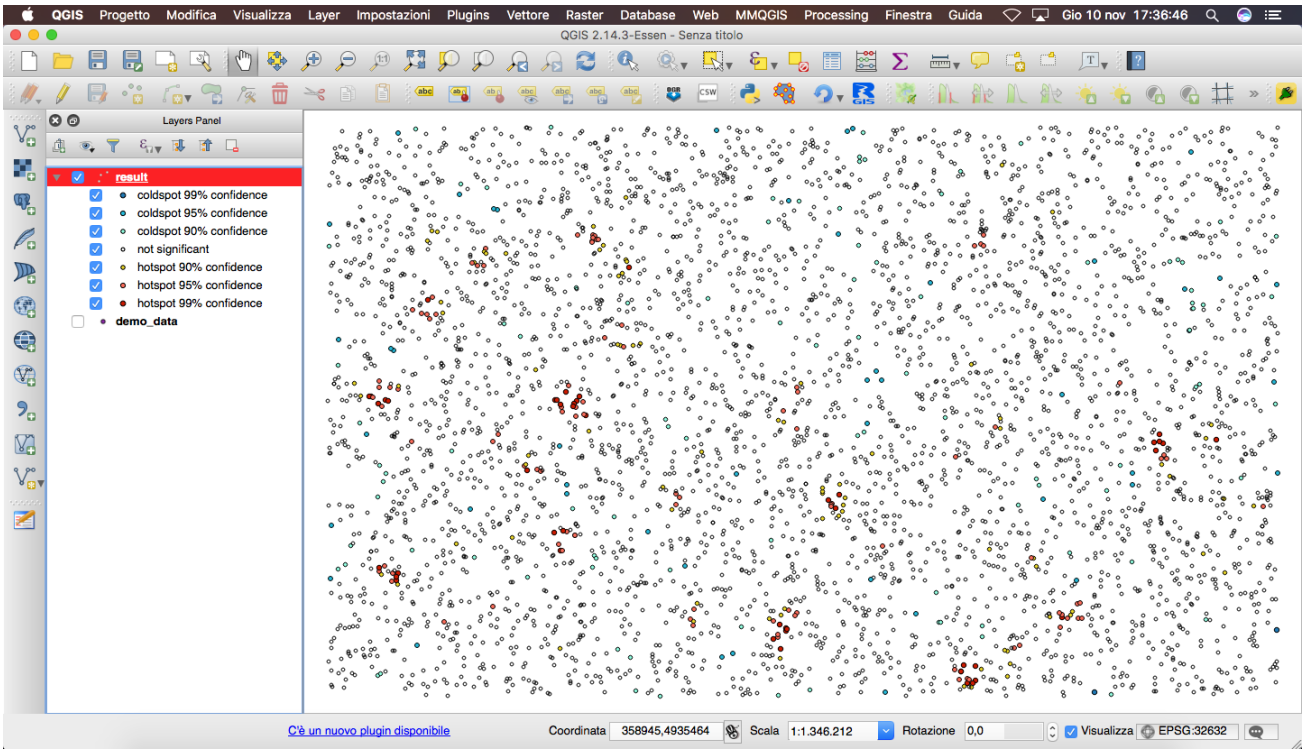


3. Plugin run and interpretation of the results

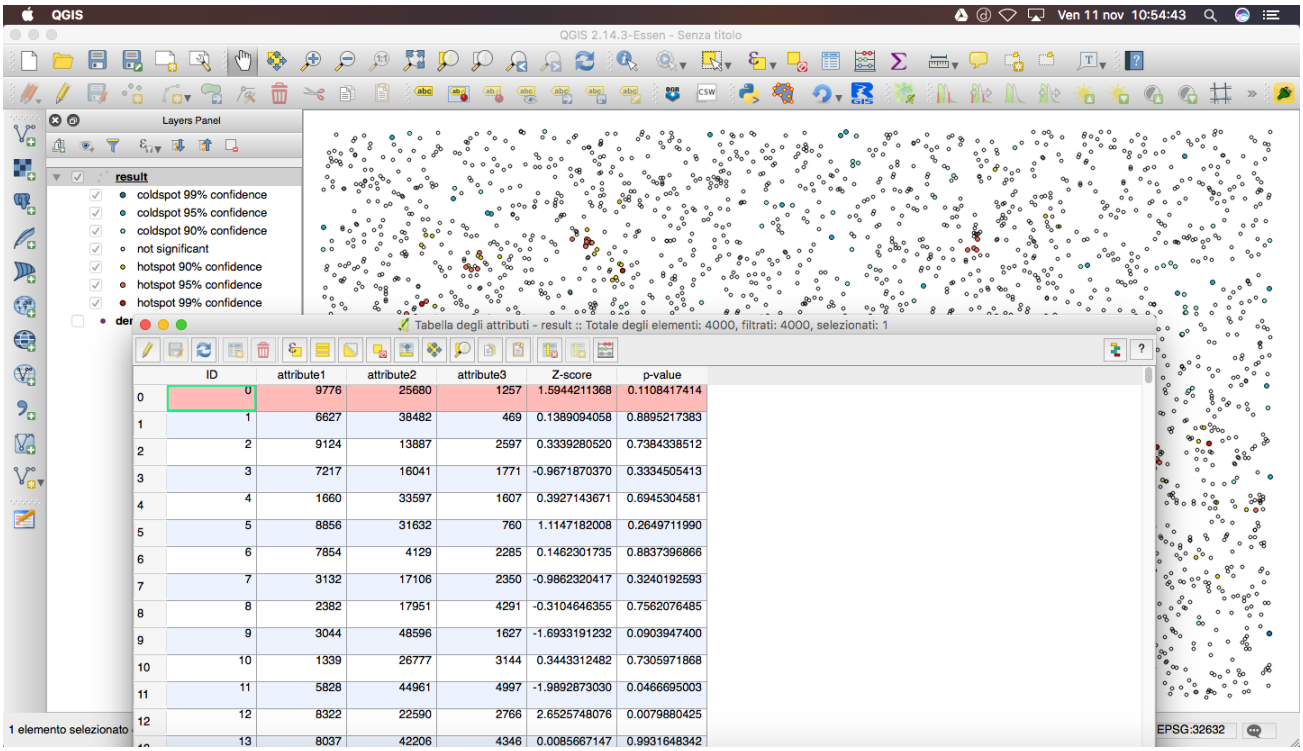
Once all the parameters are set and you have specified the output file name and path, you are ready

to press **ok** and run the Hotspot Analysis!
At the end of the process (if everything was ok) a **success message** will be displayed and the output layer will be added automatically to the QGIS map panel. The success message includes also a reminder of the Distance Band you select. If the “Optimize Distance Band” option has been used, the distance reported is the optimized one that the algorithm selected for you. The generation of the output may take some time depending on the parameter you set as discussed before.



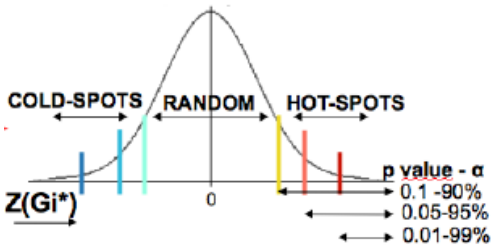


To the output layer is assign an automatic style which enable to distinguish between hotspots a coldspots, by accounting also for their statistical significance.
Let's have a look to the attribute of the output layer (in the example this layer is called "result")



	ID	attribute1	attribute2	attribute3	Z-score	p-value
0	0	9776	25680	1257	1.5944211368	0.1108417414
1	1	6627	38482	469	0.1389094058	0.8895217383
2	2	9124	13887	2597	0.3339280520	0.7384338512
3	3	7217	16041	1771	-0.9671870370	0.3334505413
4	4	1660	33597	1607	0.3927143671	0.6945304581
5	5	8856	31632	760	1.1147182008	0.2649711990
6	6	7854	4129	2285	0.1462301735	0.8837396866
7	7	3132	17106	2350	-0.9862320417	0.3240192593
8	8	2382	17951	4291	-0.3104646355	0.7562076485
9	9	3044	48596	1627	-1.6933191232	0.0903947400
10	10	1339	26777	3144	0.3443312482	0.7305971868
11	11	5828	44961	4997	-1.9892873030	0.0466695003
12	12	8322	22590	2766	2.6525748076	0.0079880425
13	13	8037	42206	4346	0.0085667147	0.9931648342

As you it can be seen in the picture, the output layer is a copy of the input layer with two new fields in the attribute table. These contains Z-scores of Gi* and related p-values at any point of the dataset. The default style of the output layer uses combination of Z-score and p-values to distinguish between hotspots and coldspots while displaying their statistical significance. Threshold values are associated to the normal standard distribution as show in the following picture. Output styling can be changed according to user's needs.



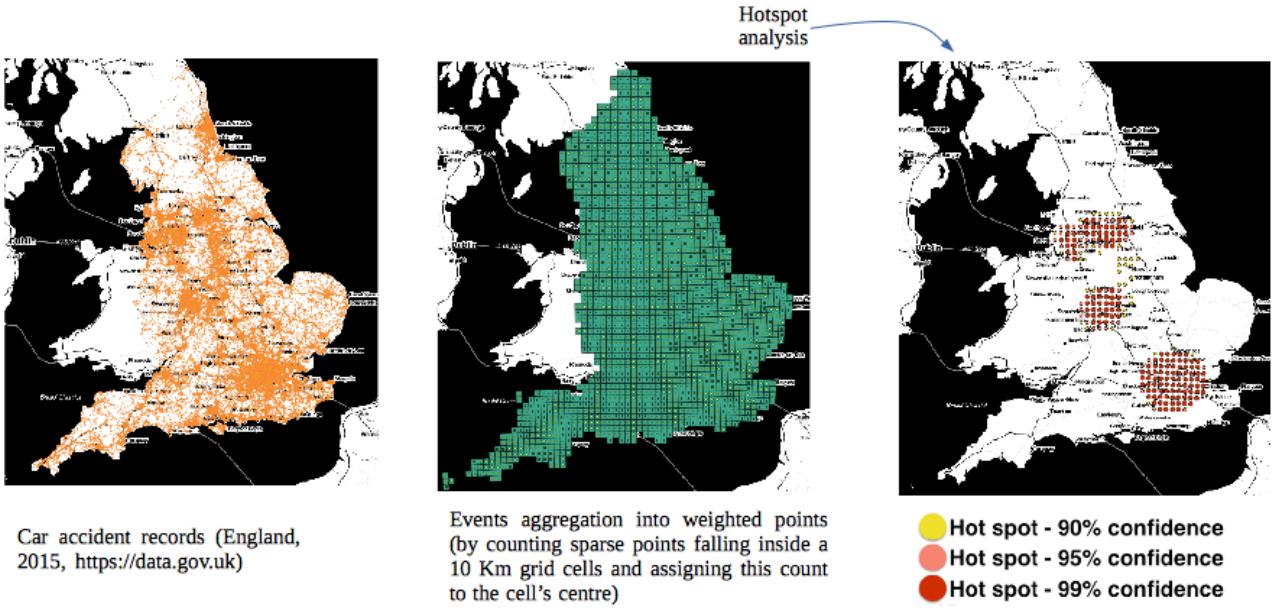
coldspot 99% confidence	"Z-score" <= -2.58 AND "p-value" <= 0.01
coldspot 95% confidence	"Z-score" <= 1.96 AND "Z-score" > -2.58 AND "p-value" <= 0.05 AND "p-value" > 0.01
coldspot 90% confidence	"Z-score" <= -1.65 AND "Z-score" > -1.96 AND "p-value" <= 0.1 AND "p-value" > 0.05
not significant	"Z-score" > -1.65 AND "Z-score" < 1.65 AND "p-value" > 0.1
hotspot 90% confidence	"Z-score" >= 1.65 AND "Z-score" < 1.96 AND "p-value" <= 0.1 AND "p-value" > 0.05
hotspot 95% confidence	"Z-score" >= 1.96 AND "Z-score" < 2.58 AND "p-value" <= 0.05 AND "p-value" > 0.01
hotspot 99% confidence	"Z-score" >= 2.58 AND "p-value" <= 0.01

Hotspots represent atypical high-value location surrounded by other high-value location as well (and coldspots vice-versa). Not significant points represent location in which local values are likely random distributed and so no significant clusters are there located.

This process enables to describe and visualize spatial distributions by highlighting atypical locations which can be fundamental to describe hidden patterns of your dataset.

4. Other examples – sparse point aggregation

Often users are interested to discover where a "sparse events" dataset show hotspots. For sparse events we intend "poor" points characterized by only a couple of coordinates without any numerical attribute associated. This is the case of GPS waypoints, Tweets, car accident reports etc.
In order to perform Hotspot Analysis on such datasets it is necessary to snap points together to create aggregated points (or "weighted points") and associate as attribute the number of sparse points which any weighted point represents.
This can be done, for example, by creating a multi-polygon shapefile or a vector grid covering your area of study and run the **Points in polygon** tool of QGIS. This associates to any polygon/cell the amount of sparse points within their area as a new attribute. At this point, it is possible to create a point shapefile using the **Centroids** tool of QGIS. As a result, you will obtain a point shapefile with an associated positive numerical attribute and you will be able to run Hotspot Analysis on it.



Car accident records (England, 2015, https://data.gov.uk)

Events aggregation into weighted points (by counting sparse points falling inside a 10 Km grid cells and assigning this count to the cell's centre)

In the example, the aggregation was performed on a polygonal grid and the cell centres layer was then created. The minimum Fixed Distance Band in this case was 10000 m (i.e. the cell centres distance). To include the 8 neighbour cells (also the one on the diagonal) this minimum distance has been multiplied by SQRT(2), so the Distance Band used was ~ 14,200 m.
This was only an example of application when the initial dataset consists of sparse events.
Future plugin development will focus both on automatic sparse points aggregation utilities as well as on the application of Hotspot Analysis using as input multi-polygon shapefile directly.

Daniele Oxoli
Ph. D Student, Politecnico di Milano
Polo Territoriale di Como, **GEOLab**
Email: daniele.oxoli@polimi.it
11/09/2016